Statistics and Probability Letters 144 (2019) 37–43

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Synthetically lagged models

Wesley S. Burr [a,*], Hwashin H. Shin [b,c], Glen Takahara [c]

[a] Department of Mathematics, Trent University: Gzowski College, 1600 West Bank Drive, Peterborough, ON K9L0G2, Canada
[b] Population Studies Division, EHRSB, Health Canada: Ottawa, ON, Canada
[c] Department of Mathematics & Statistics, Queen's University: Jeffrey Hall, 48 University Ave, Kingston, ON K7L 3N6, Canada

ARTICLE INFO

Article history:
Available online 17 July 2018

Keywords:
Air pollution
Generalized additive models
Lag
Phase
Multitaper spectrum estimation

ABSTRACT

Epidemiologic studies often report health risk due to short-term air pollution exposure through examination of *lagged* association, expressing manifestation of population health effects over time. We develop a new approach to the problem of lagged manifestation of air pollution exposure: *synthetically lagged* models. This approach also generalizes to other lagged time series regression models.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

A number of studies have shown that short-term exposure to air pollutants is associated with adverse health effects on population, in particular increased risk of premature mortality, Dominici et al. (2000). The short-term adverse health effects of air pollutants (short-term risks) are typically estimated using Poisson Generalized Linear Models (GLMs) or Generalized Additive Models (GAMs). Extending beyond univariate primary predictor models, it has become clear that same-day exposure to air pollution can have effects which only become apparent in the days, weeks, or even months following. Significant effort was expended in understanding the distributed lag structure of the air pollution-health association through classic distributed lag models (Schwartz, 2000; Roberts, 2005). This was followed by the development of more sophisticated frameworks for distributed lag, Welty et al. (2009) and Gasparrini et al. (2010).

In this work we propose a new approach for estimation of distributed lag risk through use of a method we refer to as *synthetic lag*. This method has an intuitively appealing statistical framework that allows for selection of significant structural components within a pollutant series, which we suggest are responsible for driving the risk estimates, and lagging these components. Recombining the lagged components into a single synthetically lagged pollutant series, risk estimates can then be obtained using GAMs or GLMs without the use of additional frameworks.

2. Models for estimation of population health risk due to air pollution

Research by Schwartz et al. (1996), Dominici et al. (2000) and Dominici et al. (2002) (and others), led to the development of the standard model for estimation of population health risk due to air pollution. Built around a Poisson GLM or GAM, the logarithm of daily mean mortality counts is modelled as an additive combination of a smooth function of daily mean temperature, at least one smooth function of time, a categorical variable for the day-of-week effect, and at least one

* Corresponding author.
E-mail addresses: wesleyburr@trentu.ca (W.S. Burr), hwashin.shin@canada.ca (H.H. Shin), takahara@queensu.ca (G. Takahara).

https://doi.org/10.1016/j.spl.2018.07.008
0167-7152/© 2018 Elsevier B.V. All rights reserved.

parametric air pollution covariate such as PM_{10} or Ozone. A generalized version of such a model, for $\mu_t = \mathbf{E}[Y_t]$ the time-varying daily mean mortality and Y_t a daily mortality count series indexed by day, is

$$\log(\mu_t) = \beta_0 + p_t\beta_1 + \text{DOW}_t\beta_2 + \sum_{j=1}^{k} f_j(z^j, \lambda_j),$$

with p_t the pollutant of interest, measured via a daily metric, DOW_t an indicator function for the day-of-week, and the $f_j(z^j, \lambda_j)$ are k smooth functions of environmental covariates z^j with respective degrees-of-freedom λ_j (possibly including a time variate). At minimum, a typical model will include a smooth function of same-day or averaged recent-day temperature and a smooth function of time, chosen as either a non-penalized fixed-*df* cubic regression splines (due to the findings of Ramsay et al., 2003), a penalized spline, or a Slepian (SLP) (Burr et al., 2015) smoother.

Using these models, it became clear that selection of the lag of the pollutant predictor led to highly varying estimates of association. This led to 'rules of thumb' for different pollutants, e.g., Ozone has maximized association at 1-day lag. As there is no physiological constraint which should lead to integer day-lags for maximized association between pollutant exposure and health effects, these results were curious. In particular, note that the health effect counts used in typical models are aggregated across a day as pure counts, occurring at random times through a 24-hour period. Exposures for any individuals included in the daily counts will vary up to 24 h across individual, based on occurrence of the health effect. Thus, there is already an internal lagging effect which is being 'smeared' (in a spectral sense) across the counts, in addition to any such effect which comes from the daily metric of the air pollutant.

Attempting to accommodate the multiple lagged associations between air pollutant series and health effect series led researchers to models which included multiple lags of a pollutant in a single model: distributed-lag models (DLMs, Roberts, 2005) and distributed-lag non-linear models (DLNMs, Gasparrini et al., 2010). These two families, DLMs and DLNMs, are the two well-understood methods for including lag structure in models for air pollution risk estimation. In this paper, we present a new, third approach which offers some flexibility and interpretability in presentation of predictor-response relationship, at the cost of computational complexity.

3. Synthetically lagging predictors

Motivation. It is curious how the highest (by lag choice) association between air pollutants and health effect predictors varies with pollutant choice. As inspiration, as health effect predictors are typically aggregate counts over an entire calendar day, with associated variation in individual exposure lag, how is it that ozone consistently shows highest association at one-day lag?

The sub-field of time series analysis commonly called *spectral analysis* provides several tools for examination of lag structure between two time series: in this area, lag is referred to as *phase*. For the following, an *Online Supplement* has been prepared which expands on the fundamentals, with sub-sections referred to as, e.g.,§1. Given two time series, estimation of their individual spectral densities gives complex functions of frequency; the modulus whose square is the more commonly seen real *power spectrum* and the argument, referred to as the *phase* (§1). Estimation of the *cross-spectrum* gives similar modulus and argument, with a phase plot now showing the relative phase (or lag) between the two series. Thus, it is possible to observe lag on a much finer resolution than previously considered, by simply moving to the frequency domain and considering the cross-spectrum of the two series. Coherency is the frequency-specific analogue of correlation, and can be derived from these (§3).

An example of such an exploratory lag analysis is demonstrated for Toronto, Ontario (Fig. 1 using daily maximum 8-hour mean ozone concentration and daily all-cause mortality. In this figure, each circle represents a discrete frequency, with the filled circles representing *coherent* (at 90% significance, §3.1) structure between the two series, and the coloured lines corresponding to the coarse day-at-a-time lagging which is possible in the classic models. As in most time series analyses, a large portion of the association between the two series is driven by the coherent portions. From the phase perspective, two series which are $\pm 180^\circ$ out of phase at a particular frequency can be considered to be positively/negatively associated at the lag corresponding to the frequency.

Examining Fig. 1, we note three interesting features: (1) a cluster of coherent frequencies around 365 days which appear to have a point of interest at -150°; (2) a number of clusters of coherent frequencies between 0.5 and 1.5 μHz which roughly line up with the 1-day lag line; and, (3) the majority of the coherent frequencies lie between the ± 1 day lag lines. For (1), both series (ozone and mortality) have strong yearly periodic structures, but have peaks/troughs at different parts of the year. As ozone is highest in the summer and mortality in the winter, this tracks exactly: they are approximately 5-to-6 months offset from one another, half of the yearly period of 365 days.[1] For (2) and (3), the clusters of points are largely what drives the commonly seen one-day maximum association between ozone and mortality: coherent structure with phase aligned with a one-day lag increases the regression association obtained.

Thus, we have motivation for interpreting the lagging of predictors in a time series regression model as coarsely sliding the phase-alignment of the two series around in integer-day-blocks, and by doing so, attempting to capture some portion of the underlying periodic phase structure. As we can examine this phase structure directly, §1–3, it should be possible to instead capture the behaviour of individual periodic components directly.

[1] If we reversed the order of the series in the computation of the phase, these phases would reverse and become clustered around $+150^\circ$.

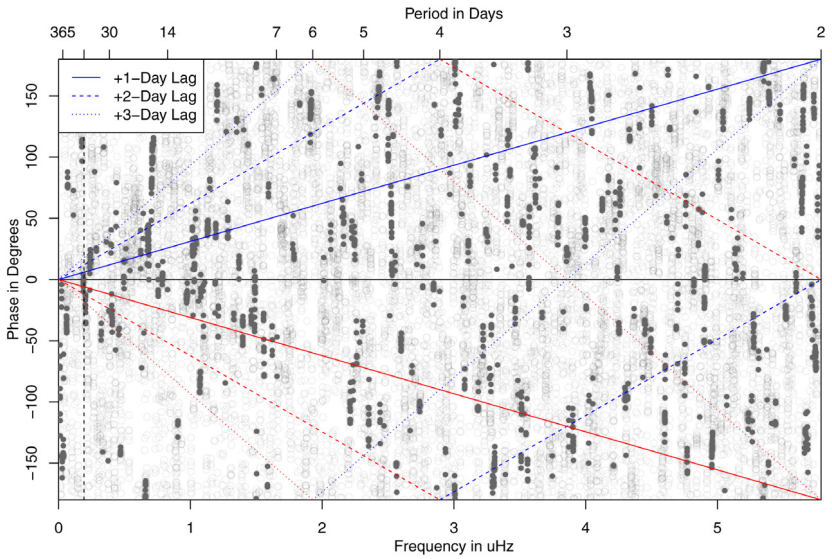

Fig. 1. Phase between daily maximum 8-hour mean ozone concentration and all-cause mortality for Toronto, Ontario, 1981–2009. All phases are plotted in light grey, with phases corresponding to 90% coherent structure overplotted as dark grey filled circles. The red lines are the negative lags of the matching positive blue lags. The black vertical dashed line at 60 days period represents the typical break-point for acute exposure models.

Lagging periodic components. When we lag an input in a time series-regression model by an integer multiple of the calendar day, we are effectively lagging individual contributory components within the series by different amounts. What we desire is to instead lag individual components by specific, targeted amounts, so as to phase-align the components and the response series. For the purposes of this approach, we will consider a time series x_t to consist of the sum of a stationary random series Z_t plus M periodic components

$$x_t = Z_t + \sum_{j=1}^{M} \alpha_j \cdot \cos(\omega_j t + \phi_j),$$

each with amplitude α_j, frequency ω_j (in radians per unit time) and phase ϕ_j. This is the classic discrete- and finite-time spectral representation. As this can apply to both the predictor and response in a time-series regression model, classical theory tells us that cosinusoids at different frequencies are uncorrelated, implying that the association between predictor and response can be modelled as arising from both common frequencies and the association (correlation) between their respective stationary random processes, $Z_t^{(1)}$ and $Z_t^{(2)}$.

Synthetic lag algorithm. To lag individual periodic components requires two elements: the estimation of the phase offset at a particular frequency, and the implementation of the individual offsetting action. To estimate the phase, we use multitaper spectrum estimation (Thomson, 1982) (§1–2), and estimate the multitaper coherence (§3) between the two series. This gives a complex estimate of coherence consisting of a magnitude and a phase. The phase estimated via this approach is a robust estimate of the frequency-specific phase between the two series (Chave et al., 1987). As the trade-off of the multitaper is the reduction of broad-band bias for local bias, the phase must be estimated across a local bandwidth in the neighbourhood of a coherent result. Note that a probabilistic clip (e.g., 95%) is a required argument for the algorithm, designed to allow the phase-alignment procedure to be applied to domain-specified coherent structure. As the extended simulation in §4 shows, decreasing this argument will, in general, increase the net association of the resultant series through inclusion of more candidate frequencies for alignment.

The second element of the algorithm is the implementation of the offset. The Fourier transform, $\mathcal{F}(x_t)$, of a time series results in a *complex* function of frequency. As such each frequency consists of a complex modulus and argument, §2. Multiplying such a complex result by a complex number of modulus one and argument corresponding to the phase-offset rotates the complex frequency-specific result so as to "line up" the two series in phase. The resulting complex transform result can then be inverted back into the time domain to obtain a phase-aligned version of the original series. That is,

$$X(f) = \mathcal{F}(x_t) \qquad \text{and} \qquad Y(f) = \mathcal{F}(y_t)$$
$$X(f_0 - W, f_0 + W) = X(f_0 - W, f_0 + W) \cdot e^{i\phi_{\text{off}}}$$
$$\hat{x}_t = \mathcal{F}^{-1}(X)$$

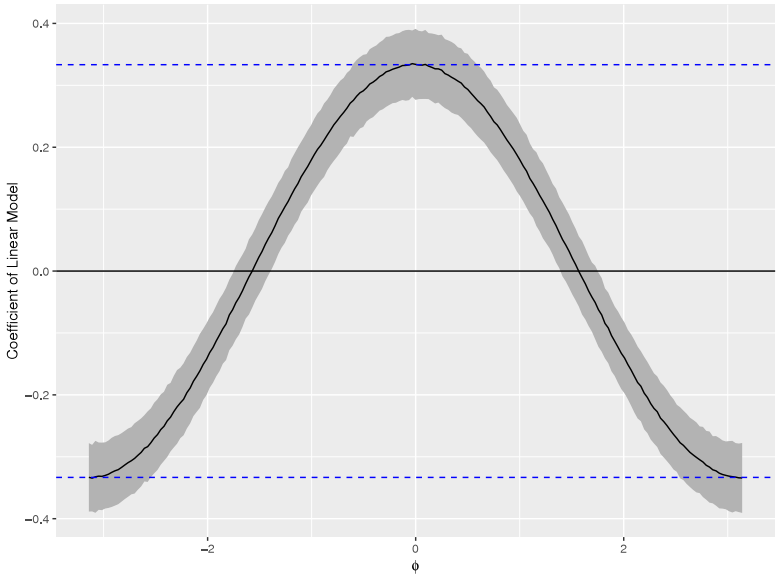

Fig. 2. Simulation 1, showing changes in the coefficient of the predictor in a linear model, $y \sim x$, across changing values of ϕ. The ribbon represents a 95% confidence interval for the estimate of the coefficient, computed across 1000 realizations of each simulation ϕ-step.

In this formulation, ϕ_{off} is measured as the phase difference between the response and predictor for the given frequency, so that rotating the predictor by ϕ_{off} will result in approximately equal phases. The actual implementation can be made arbitrarily complicated by consideration of subtleties in each stage of the algorithm (estimation of the complex eigencoefficients, estimation of the complex phase offset, bandwidth used for estimation of the phase offset, application of the phase offset to the series, etc.). An implemented version of this algorithm in code is available (Burr, 2018). The implementation uses a half-W bandwidth (Thomson, 1982) to allow for departure from pure cosinusoid nature and applies phase offsets frequency bin-at-a-time (with each bin receiving its own estimated phase application).

4. Simulations

All simulations from this section have been made available via Burr (2018) along with the implemented synthetic lag algorithm. Consider two time series, each with one periodic component $\alpha^{(j)}\cos(\omega^{(j)}t+\phi^{(j)}), j = 1{,}2$ and each with a stationary random series $Z_t^{(j)}$. For Simulation 1, we assume that each $Z_t^{(j)} \sim \text{WN}(0, \sigma^2)$, e.g., no correlation between $Z_t^{(1)}$ and $Z_t^{(2)}$. We set x_1 to be the predictor and x_2 the response, $t \in \{1, 2, \cdots, 1000\}$, $\alpha^{(1)} = \alpha^{(2)} = 1.0$ and $\omega^{(1)} = \omega^{(2)} = \frac{2\pi}{4}$ (4-day period). We then allow $\phi^{(1)}$ to vary smoothly from $-\pi$ to π radians while keeping $\phi^{(2)}$ fixed at 0. Since this simulation represents daily data, the time step $\Delta t = 86400$ seconds (1 day). When $\phi^{(1)} = \phi^{(2)}$, the expected covariance between the two series is 0.50 precisely, with correlation (and linear model coefficient) $1/3 = 0.333$. We compute 1000 iterations of the simulation for each $\phi^{(1)}$, and show the coefficients and their standard errors in Fig. 2 as a function of $\phi^{(1)}$.

We see that our intuition for a single periodic component matches the actual performance. When the phases of the two series are aligned, the coefficient of the linear model is the correlation coefficient between the two, and when the phases are not aligned, the coefficients decrease in systematic ways, achieving a minimum at $\phi = \pm\pi$. This represents a phase difference of 180°and corresponds to a covariance of -0.5 and a correlation of $-1/3$, as seen.

For Simulation 2 we again consider two time series, but each with two periodic components $\alpha_i^{(j)}\cos(\omega_i^{(j)}t+\phi_i^{(j)}), i, j = 1{,}2$, and each again with a stationary random series $Z_t^{(j)}$. We again assume that the $Z_t^{(j)} \sim \text{WN}(0, \sigma^2)$, and set all α terms to 1.0. The frequencies are set to $\omega_1^{(j)} = \frac{2\pi}{4}, j = 1, 2$ and $\omega_2^{(j)} = \frac{2\pi}{6}, j = 1{,}2$. The response, x_2, has $\phi_1^{(2)} = \phi_2^{(2)} = 0.0$, and the predictor, x_1, has phase terms allowed to vary independently from $-\pi$ to π. Fig. 3 shows the coefficient as a function of $\phi_1^{(1)}$ and $\phi_2^{(1)}$, with values ranging from (min) to (max). As expected, the maximum of 0.50 occurs when $\phi_1^{(1)} = \phi_2^{(1)} = 0.0$, as this perfect agreement between the two pairs of deterministic components of the two series. We also see lines of high coefficient along the zero-lines of each phase argument, and minimums at the four corners where $\phi_1, \phi_2 \in \{\pm\pi\}$.

Simulations for phase-alignment. We now proceed to examine the synthetic lag argument from the perspective of simulation. For Simulation 3, we expand the setup of Simulation 1, and generate 10 periodic cosinusoids with random amplitudes, $\alpha_j \sim U(0.5, 3)$, random frequencies, $f_j \in 2\pi \cdot \{1, 2, \ldots, 99, 100\}^{-1}$, and random phases, $\phi_j \in [-\pi, \pi]$. We restrict the response series to have no phase offsets, but equal frequencies and amplitudes to the predictor series. Our baseline is the estimation of correlation between the two series with no phase offsets applied, and each iteration of the simulation

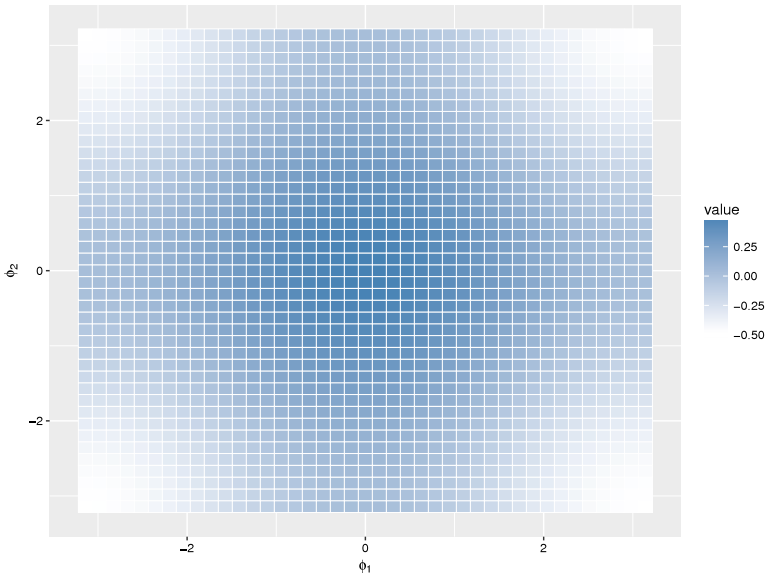

Fig. 3. Simulation 2, showing changes in the coefficient of the predictor in a linear model, $y \sim x$, across changing values of the two phase arguments of x_t, ϕ_1 and ϕ_2. The maximum occurs at $\phi_1 = \phi_2 = 0$ and is 0.5, while the minimum occurs in all four corners and is -0.5.

Table 1
The left table shows $M = 250$ realizations of a simulation with 10 periodic components with equal amplitudes and frequencies, but different phases. Results reported as percentage of the true correlation between the two series upon perfect phase-alignment. The right table shows the realization-by-realization comparison of performance between the Synthetic Lag procedure and integer-lags. Columns are the empirical percentiles of the results.

	2.5%	50%	97.5%	SynthLag -	2.5%	50%	97.5%
Lag-0	−77.12	1.21	62.12	Lag-0	11.09	44.50	86.63
Lag-1	−76.03	−0.75	64.98	Lag-1	12.13	44.25	88.75
Lag-2	−74.11	−2.80	67.49	Lag-2	10.32	44.87	92.31
Lag-3	−75.82	−3.05	66.44	Lag-3	6.16	45.74	92.26
SynthLag	−23.08	46.76	97.87				

is measured as percentage of the true correlation, with sign preserved. We iterate for $M = 250$ realizations with $N = 1000$ series length, and a cutoff of 0.90 for the synthetic lag procedure. (§4 has an exploration of these parameter choices.) The results are shown in Table 1. Note that generation of null data gives the expected null result, §5.

The left-hand table of Table 1 shows that the synthetic lag procedure attains up to 98% of the true correlation as a maximum, and −23% of the true correlation as a minimum. This implies that there are some cases for which the sign of the result remains incorrect (negative rather than positive). However, the right-hand table of Table 1 shows the realization-by-realization empirical percentiles, and shows that there is not a single realization where the synthetic lag procedure does not improve (in the sense of increasing toward the true correlation) on any of the chosen integer-valued lags.

5. Real-world data

We now transition to a real-world data set: the NMMAPS database (Dominici et al., 2000), using one of the most complete records in the database: Chicago, Il., from 1987 to 2000 (5114 days). We model response `Death` (All-Cause Mortality) against pollutant PM_{10} and account for same-day temperature, day-of-week, and including a smooth function of time. We use Slepian `slp` smoothers for the smooth function of time, Burr et al. (2015). The model used is a simplified version of that of Dominici et al. (2000), and is compared to the same model with the PM_{10} lagged pollutant replaced with a synthetically lagged version of the starting lag, e.g., lag-2 PM_{10} has the synthetic lag algorithm applied to it, and the model is then estimated. The associations of the pollutant predictor are shown in Fig. 4.

As the results show, application of the synthetic lag algorithm results in an increase of the estimated association in the direction of the maximum (true) association, under the (reasonable) assumption that all coherent structure is related. Each of the associations in Fig. 4 starts from a different base series: the discrete day-lagged versions of PM_{10}, lags 1 through 4. The changes in the series due to the synthetic lag algorithm are different for the respective day-lagged versions, as the association of the underlying noise series, Z_t, is different. Note that these results are statistically different: the confidence intervals shown are 95% intervals under assumption of normality, and are clearly delineated.

In Fig. 5 we examine the comparison of the aggregate relative risk on Cardiovascular death (code `cvd`) due to Ozone, using both the DLNM (Gasparrini et al., 2010) and our new synthetic lag algorithm, across the largest cities in the United States.

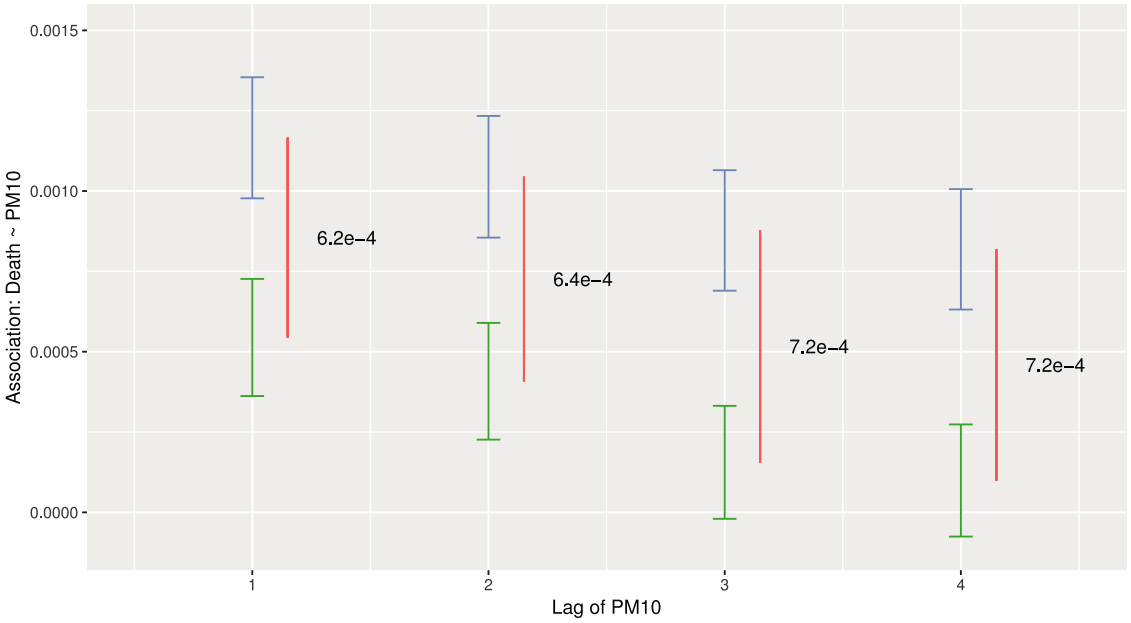

Fig. 4. Associations between variables `Death` and `pm10tmean` for lags 1 to 4, after accounting for temperature, day of week, and smooth function of time. The larger (blue) CIs are the synthetic lag results, while the smaller (green) CIs are the single-lag predictor results. The red vertical lines are a visual aid to indicate differences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

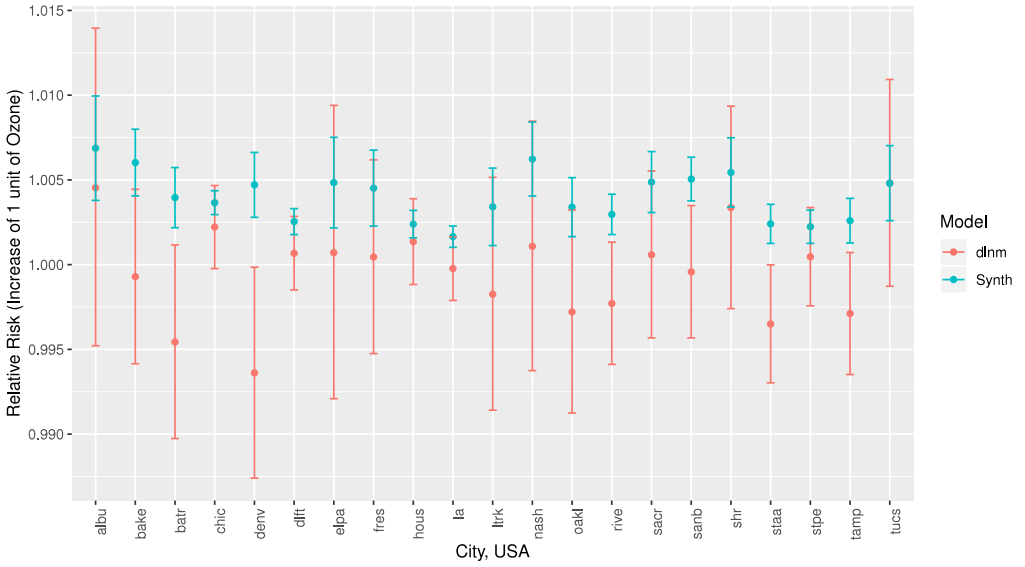

Fig. 5. Application of the DLNM (red) and synthetic lag (turquoise) algorithms to a selection of large population US cities for Ozone as predictor and CVD as response. Results reported as relative risks. As in the previous example, these associations are after accounting for temperature, day-of-week, and a smooth function of time. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The DLNM algorithm aggregates risk from multiple lags together in an identifiable way, but as it is restricted to discrete day-length lags, the synthetic approach identifies and aligns structural components that are not captured by the former, and results in quite different, and consistently more positive, results.

The logic underlying the three methods compared indicate that the phase alignment at the heart of the synthetic lag approach is capturing association missed by the other algorithms. As the entire procedure can be completed before analysis in the context of other predictors or before accounting for any other component, it allows for comparison to most previously developed algorithms.

6. Discussion and conclusion

We have developed a new algorithm for the identification and lagging of periodic structural components in time series regression models. The synthetic lag algorithm allows for the equivalence of using multiple discrete integer lags, without the inherent issues involved, namely: correlation of the underlying noise sequence, identifiability of the coefficients obtained, and in-model computational complexity. The trade-offs are the requirements for a contiguous series (to transform to the frequency domain and estimate coherence and phase) and the increased computational complexity from the pre-model stage.

This method has further applications in a wide variety of fields, essentially any domain in which lagged time series regression models are used, including infectious disease forecasting models, geophysical and electromagnetic models, and the example used in this paper — environmental epidemiological models. We hope to extend this new algorithm to the consideration of the inter-predictor complexity inherent in multiple predictor-models, in particular two- and three-pollutant time series regression models for air pollution and health research. In addition, further development allowing for the propagation of the associated uncertainty in frequency and phase estimation to the final model outputs would allow for increased precision and interpretation of the final model outputs.

Acknowledgements and Funding

We thank Drs. Rick Burnett (Health Canada) and David J. Thomson (Queen's University) for feedback on earlier versions of this work. Funding was provided by Health Canada via grant programs 'CESI' (859104) and 'CARA' (4500335886) and by a CANSSI collaborative team research project, "Modern Spectrum Estimation in Time Series Analysis".

Appendix A. Supplementary data

Supplementary explanatory material related to this article can be found (online https://doi.org/10.1016/j.spl.2018.07.008).

References

Burr, W.S., 2018. Code for 'Synthetic Lag Models', GitHub Repository, https://github.com/wesleyburr/synthLag. http://doi.org/10.5281/zenodo.1246898.

Burr, W.S., Takahara, G., Shin, H.H., 2015. Bias correction in estimation of public health risk attributable to short-term air pollution exposure. Environmetrics 26 (4), 298–311.

Chave, A.D., Thomson, D.J., Ander, M.E., 1987. On the robust estimation of power spectra, coherences, and transfer functions. JGR 92, 638–648.

Dominici, F., McDermott, A., Zeger, S., Samet, J., 2002. On the use of generalized additive models in time-series studies of air pollution and health. Am. J. Epidemiol. 156 (3), 193–203.

Dominici, F., Samet, J., Zeger, S., 2000. Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. JRSS:A 163 (3), 263–302.

Gasparrini, A., Armstrong, B., Kenward, M., 2010. Distributed lag non-linear models. Stat. Med. 29 (21), 2224–2234.

Ramsay, T., Burnett, R., Krewski, D., 2003. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. Epidemiology 14 (1), 18–23.

Roberts, S., 2005. An investigation of distributed lag models in the context of air pollution and mortality time series analysis. J. Air Waste Manage. Assoc. 55 (3), 273–282.

Schwartz, J., 2000. The distributed lag between air pollution and daily deaths. Epidemiology 11 (3), 320–326.

Schwartz, J., Spix, C., Touloumi, G., Bachárova, L., et al., 1996. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. J. Epidemiol. Community Health 50 (Suppl 1), S3–11.

Thomson, D., 1982. Spectrum estimation and harmonic analysis. Proc. IEEE 70, 1055–1096.

Welty, L.J., Peng, R., Zeger, S., Dominici, F., 2009. Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. Biometrics 65 (1), 282–291.